The BBN RT03 BN Arabic System

Long Nguyen,
John Makhoul, Mohamed Noamany

RT-03 Workshop

Boston, MA, May 19-21, 2003

Overview

- Development test set
- Improvements
- Evaluation result

BBN TECHNOLOGIES
A Verizon Company

Language-Specific Treatment

- Transcripts: Used a modified Buckwalter format

  For example: 'JxbAr ywm kAml nJtykm bhA fy brnAmjnA AlMxbAry Al$Aml AlsAEp AlrAbEp wAlE$rwn…'

- 'Phonetic' dictionary: Used grapheme-phoneme, one-to-one mapping rule
- High OOV rate mainly due to affixes
  - ~5% OOV rate for a 65k-word lexicon

Development Test Set

- Selected four episodes from the two audio sources in the TDT4 Arabic corpus
  - NTV: Nile TV shows from Cairo, Egypt
  - VOA: Uncle Sam's radio show
  - Broadcast in the second half of Jan '01
  - First 30 minutes from each episode (~2 hours)

	NTV	VOA	All
Baseline (GI, 1xRT)	28.6	20.9	24.6

Improvements

- Used GD acoustic models (36h of LDC data), MLLR adaptation, and ran at 10xRT
- LM: 65k-word lexicon, 15M bigrams, 33M trigrams, trained on mostly in-house data
- Added 38h of TDT4 data

	NTV	VOA	All
0. Baseline	28.6	20.9	24.6
1. GD, adaptation, 10x	23.6	16.8	20.0
2. + TDT4 data (38h)	22.3	14.5	18.2

Evaluation Results

- System ran at 7xRT
- Evaluation result (26.3%) is significantly higher than the result on the dev test
  - Self-score using NIST's reference is 26.7%
  - There are problems in the reference
- *[Adjudication ongoing]*

	NTV	VOA	All
Dev03 (2h)	22.3	14.5	18.2
Eval03 (1h)	25.3	28.5	26.7

Summary

- Achieved good WER reduction on the development set
- Need to agree on a transcription standard for Arabic